Pathway-based microarray gene expression analysis use cases

Gary Bader, Ethan Cerami, Chris Sander

Motivation: One of the main issues with any large-scale experimental method is interpreting the meaning of the large amount of information output and prioritizing it for further study. This is particularly challenging with transcriptional profiling using microarrays that contain as many as 40,000 transcript variants representing tens of thousands of human genes. A typical transcriptional profiling experiment that compares, for instance, a tumor transcript profile to a control normal tissue transcript profile can generate a list of thousands of genes that are deemed interesting for further study because they are differentially expressed in the tumor sample. Obviously, these genes are also associated with additional information, such as expression values, fold-change values and significance statistics. One useful line of analysis to pursue on such a large gene list is functional or pathway analysis, where the expression profile is studied in the context of known biological process and molecular pathway information. The expression data is expected to be structured according to known biological processes, since it is known that biological pathways are coordinately regulated by transcriptional and translational processes. Thus a large list of differentially expressed genes can be mapped to a much smaller list of differentially expressed pathways to ease further analysis.

The main challenge with this class of analysis techniques is integration of pathway data from multiple heterogeneous data sources.

How we do pathway-based microarray gene expression analysis now.

Use case 1: Visualize gene expression data on a pathway

User wants to see their data in the context of known pathways so that they can manually inspect if a pathway is up or down-regulated or which parts of a pathway are active.

Data requirements: Pathway diagrams with computer readable information about gene position in the diagram. These are stored in various image formats, both static and dynamically generated in different databases. An example from caBIG is eMIM.

User loads their gene expression data into a visualization tool, such as GenMAPP, and views their gene expression data in the context of a pathway diagram. GenMAPP is able to map gene expression values to color genes on the diagram by expression.

Advantage: User is able to examine their expression data in the context of their own biological knowledge in a very detailed way.

Disadvantage: Process is time-consuming and significance of findings is hard to judge.

Use case 2: Pathway set based gene expression analysis

User wants to see if pre-defined pathways are significantly over-represented in user defined subsets of their gene expression data set using tools such as GOMiner. User defined subsets are usually defined by either:

1. Differentially expressed genes between two experiments, where the threshold for differential expression is chosen arbitrarily
2. A cluster of co-expressed genes generated by a clustering algorithm

Data requirements: Sets of genes that are known to be in the same pathway (or other gene sets). These can be extracted from pathway databases, such as KEGG, by discarding the connectivity information in the pathway and from resources, such as the Gene Ontology Biological Process Ontology annotation, which define which genes are part of certain pathways, but don't contain any connection information between the genes e.g. that one protein binds to another.

Advantages: Statistically significant over-representation of sets of genes in expression data is calculated. Pre-defined well-known over-represented pathways are shown to the user.

Disadvantages: Almost always heavily biased by arbitrary gene set definition.

Use case 3: Pathway- and network-based gene expression analysis

User wants to find regions of a network of molecular associations that is co-expressed given their gene expression data. These regions are expected to represent active parts of pathways.

Data requirements: Large sets of molecular interactions and annotated pathways. Data is collected using the following steps:
Step 1: Collect pathway data from major pathway data sources manually (download from various web pages and FTP sites) in various file formats.
Step 2: Extract data from various file formats using Perl to create simplified networks of associated molecules (type of association may be used in a simplified way e.g. activates/represses, regulates, binds, etc.)
Step 3: Load into an application that uses a custom or proprietary file format for analysis.

Advantages: All relevant pathway and expression data is taken into account in an unbiased manner to find pathways or regions in the interaction network that are significantly co-expressed. Multiple unrelated genes that are co-expressed by chance that would normally cluster together by gene-expression alone can be separated based on biological pathway knowledge.

Disadvantages: Molecular interaction information is far from complete, so a lot of gene expression data is not taken into account. Well-known pathways are not always identified, potentially leaving the user without known reference points in their data.

Common architecture relevant problems among use cases:

Architecture relevant issues:
The main architecture relevant issues are related to gathering enough data to be able to perform the above use cases. The more high quality pathway data that is available, the better the scientific result, thus a complete collection of all known pathway data is eventually desirable.

1. Data retrieval. Retrieving data from multiple different sources in multiple different formats is difficult. A registry of data locations as well as update cycles on the data would be useful for data integrators.

2. ID resolution: When merging data from various sources, it is very difficult to resolve gene identifiers, as every gene has many identifiers (sometimes more than 20) and each database uses subset of these identifiers (usually from 1 to 3) to identify their genes. The subset of gene identifiers used by each data source for the same gene generally does not completely overlap. Seamless and fast ID resolution (get all known synonyms given a protein/gene/small molecule ID) is necessary to merge bioinformatics data from different sources. Example web services that try to support this currently are AliasServer (http://cbi.labri.fr/outils/alias/), MatchMiner (http://discover.nci.nih.gov/matchminer/html/index.jsp) and EnsMart (http://www.ensembl.org/Multi/martview). The LSID initiative seems to provide some ID resolution services, but these seem to only be in test mode.

3. Ability to track versions and provenance of pathway data (and other input data, such as gene expression data) so that once an analysis is done, it can be repeated using the same input. This allows analysis methods and results to be compared and tracked after they are performed. A seamless version management system that used standard metadata for version and creator information that was used consistently would be needed to provide this service. This is generally ignored in typical analyses since it is such a huge problem.